NARROWING PARAMETRIC UNCERTAINTY IN LAND-HYDROLOGY MODELS

Leveraging Machine Learning Emulators for Parameter Estimation Across 124 Flux Towers Using Process-Based Models

Ignacio Aguirre[1*], Wouter Knoben[1], Nicolás Vásquez[1], Andy Wood [2], Dave Lawrence[2], Gordon Bonan[2], Will Wieder[2], Keith Oleson[2], Linnia Hawkins[3], Brian Dobbins[2], Daniel Kennedy[2] and Martyn Clark[1]

[1] Department of Civil Engineering. University of Calgary, Alberta, Canada
[2] NSF National Center for Atmospheric Research (NCAR), Boulder, Colorado, USA
[3]Columbia University, New York, USA

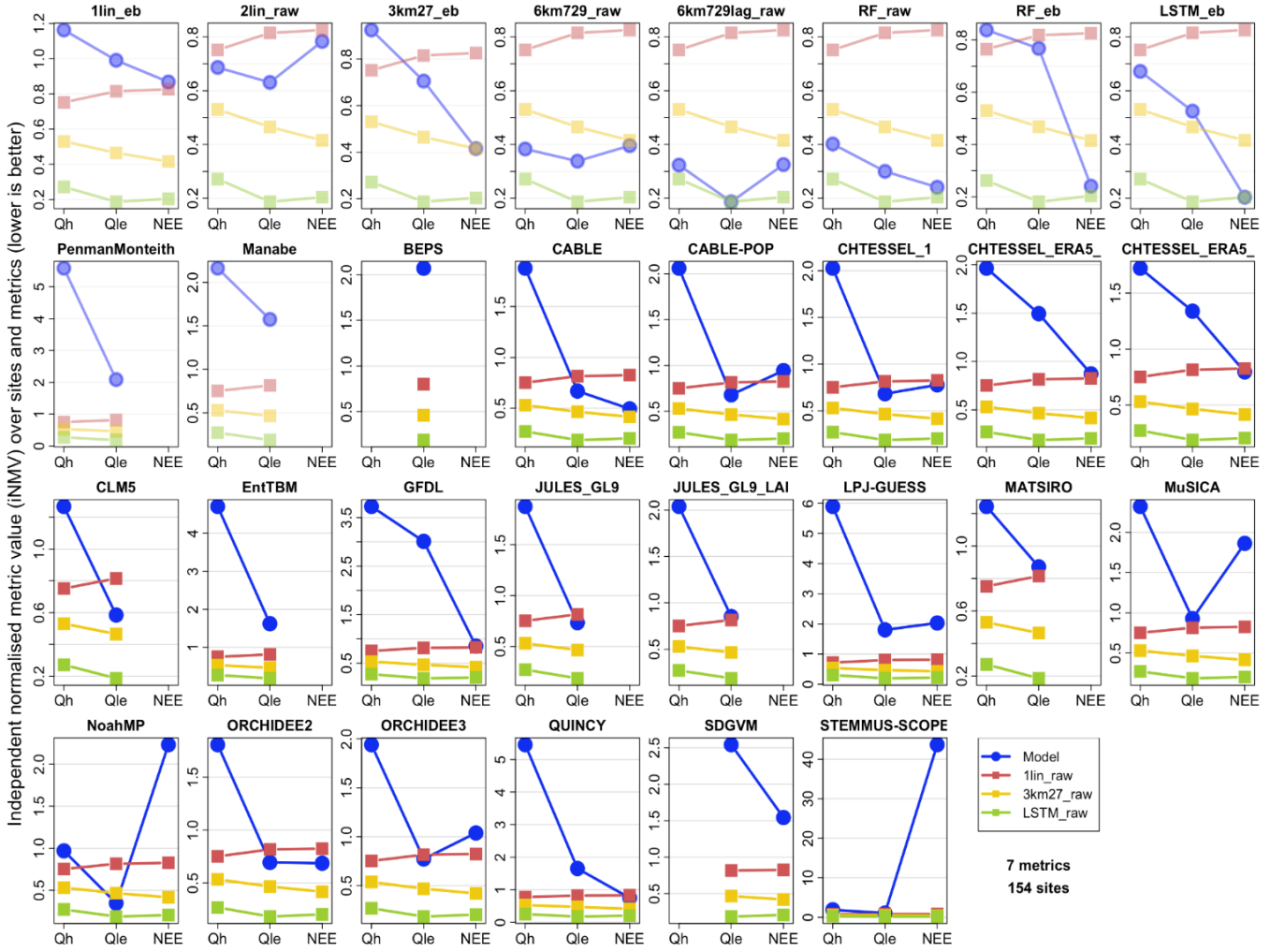
*Contact: ignacio.aguirre@ucalgary.ca

February 24, 2026
Land Model and Biogeochemistry
Winter Working Group Meetings

Introduction

Abramowitz et al. (2024) led the Plumber 2 MIP to evaluate the performance of models on turbulent fluxes (latent heat and sensible heat) over 154 flux towers. Their results show that:

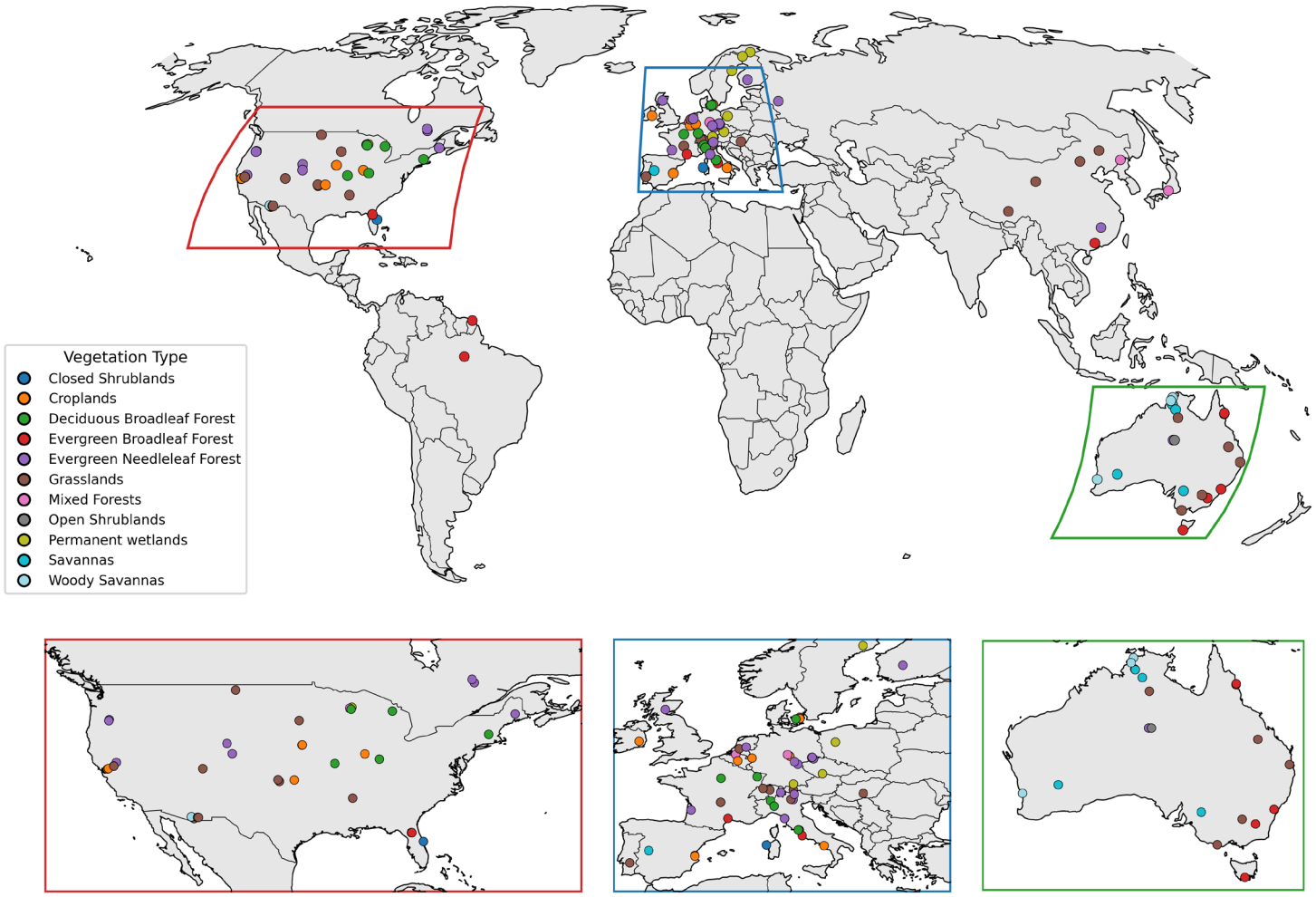
- All state-of-the-art models are outperformed by simple regression for sensible heat (Qh)
- The LSTM benchmark for latent heat (Qle) outperforms all models
- As in the catchment hydrology example by Nearing et al. (2021), they show that LSTM performs better than process-based models; **however, model performance can be improved through calibration.**



This study addresses this challenge through calibration, comparing different calibration methods (e.g., single-site emulator, large-sample emulator, genetic algorithm, and dimensional search) to identify an optimal parameter set for latent heat and sensible heat, and evaluating its performance using temporal and spatial cross-validation.

Sites

- To capture different hydrological regimes, we used the PLUMBER 2 dataset (Abramowitz et al., 2023; Ukkola et al., 2022).
- We removed stations with data issues or fewer than 2 years of data, resulting in 124 flux towers.
- We divided the data into calibration (1st 50% of the data) and validation (2nd 50% of the data)
- We only used the measured-only data for latent heat and sensible heat to compute the metrics.



Methodology

- To explore different modeling decisions, we utilized the Structure for Unifying Multiple Modeling Alternatives (SUMMA, Clark et al., 2015)
- **Simulate the conservation of mass and energy.**
- **Multiple modeling options for specific processes**
- Multiple state-of-the-art numerical solvers for the equations, including the SUNDIALS suite.
- Flexibility to adjust model parameters.
- Multiple options to represent horizontal and vertical heterogeneity.

Clark et al. (2021)

Methodology

- Using the Kling-Gupta efficiency (KGE, Kling et al., 2012) and focusing on a composed metric of 50% of latent heat (Qle) and 50% of sensible heat (Qh). We compared the default SUMMA runs against four calibration methods and two data-driven benchmarks.
- These are the calibration methods:

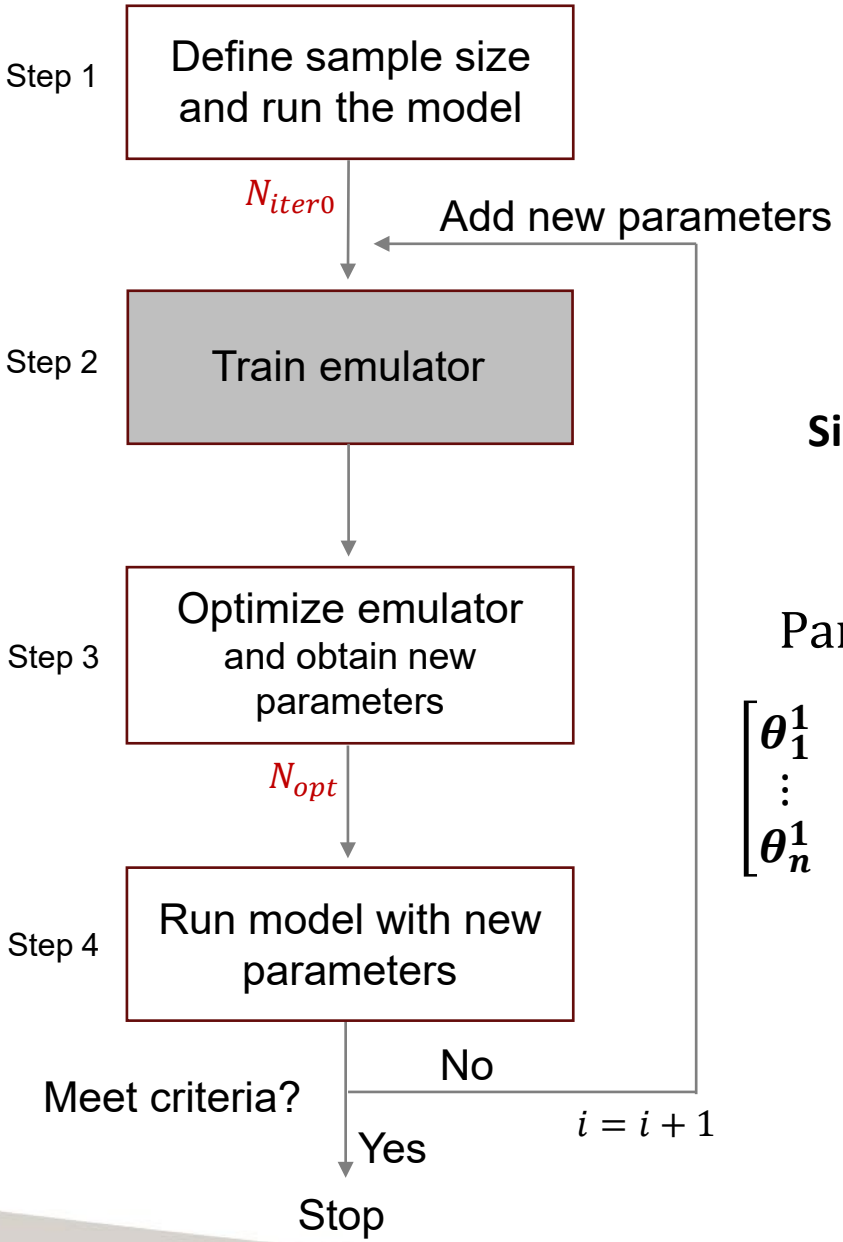
1. Calibration using the single-site emulator (SSE; Tang et al., 2025), where a machine-learning emulator is trained only with data from one site. Total runs: 2300
2. Calibration using the large-sample emulator (LSE; Tang et al., 2025), where a machine-learning emulator is trained using the data from the 124 flux towers at the same time. Total runs: 2600
3. Calibration using the Dynamically Dimensioned Search algorithm (DDS; Tolson and Shoemaker, 2007). Total runs: 2000.
4. Calibration using a genetic algorithm (GA; Yoon and Shoemaker, 2001). Total runs 2000

- Plus, we used two benchmarks based on Gauch et al. (2021) and Kratzert et al. (2022).

1. Lower-bound benchmark: a multifrequency (daily and 30-minute) MTS-LSTM (Gauch et al., 2021; Kratzert et al., 2022) trained using data from a single site. Key hyperparameters include an output dropout of 0.4, 100 training epochs, and a hidden size of 64.
2. Upper-bound benchmark: a multifrequency (daily and 30-minute) MTS-LSTM (Gauch et al., 2021; Kratzert et al., 2022) trained jointly on data from 124 sites. Key hyperparameters include an output dropout of 0.4, 30 training epochs, and a hidden size of 64

Calibration: methodology

Based on the work of Tang et al. (2025)



Single site emulator

$$\boldsymbol{\theta} \quad \boldsymbol{y}$$

Parameters Metric

$$\begin{bmatrix} \boldsymbol{\theta_1^1} & \cdots & \boldsymbol{\theta_1^P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\theta_n^1} & \cdots & \boldsymbol{\theta_n^P} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{y_1} \\ \vdots \\ \boldsymbol{y_n} \end{bmatrix}$$

$$NKGE = \frac{KGE}{2 - KGE}$$

Large sample emulator

$$\theta \quad A \quad y$$

Parameters Attributes Metric

$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_1^P \\ \vdots & \ddots & \vdots \\ \theta_n^1 & \cdots & \theta_n^P \end{bmatrix} \begin{bmatrix} a_1^1 & \cdots & a_1^A \\ \vdots & \ddots & \vdots \\ a_1^1 & \cdots & a_1^A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$ Tower 1

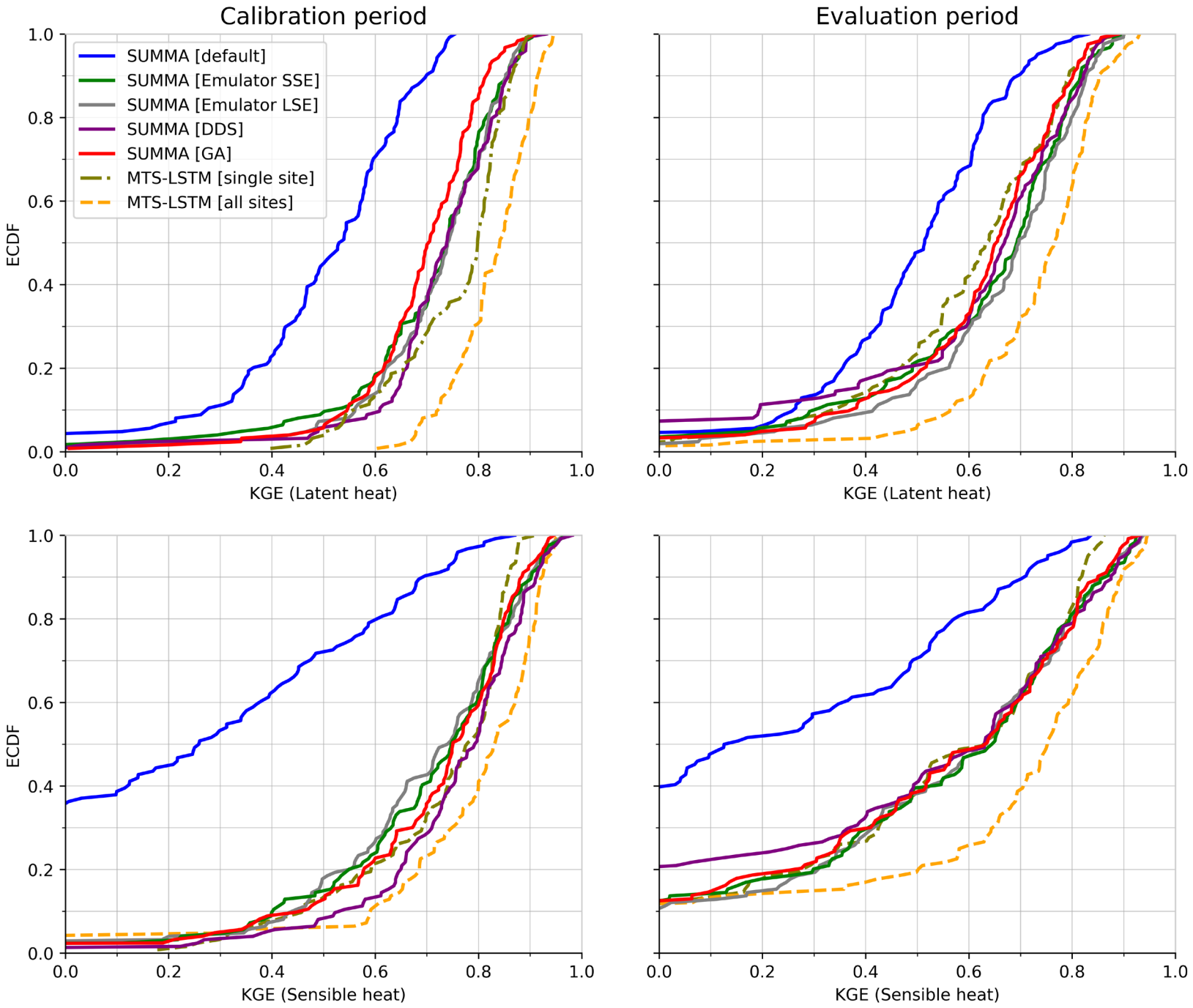
$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_1^P \\ \vdots & \ddots & \vdots \\ \theta_n^1 & \cdots & \theta_n^P \end{bmatrix} \begin{bmatrix} a_2^1 & \cdots & a_2^A \\ \vdots & \ddots & \vdots \\ a_2^1 & \cdots & a_2^A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$ Tower 2

$$\vdots \qquad \vdots \qquad \vdots$$

$$\begin{bmatrix} \theta_1^1 & \cdots & \theta_1^P \\ \vdots & \ddots & \vdots \\ \theta_n^1 & \cdots & \theta_n^P \end{bmatrix} \begin{bmatrix} a_B^1 & \cdots & a_B^A \\ \vdots & \ddots & \vdots \\ a_B^1 & \cdots & a_B^A \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$ Tower B

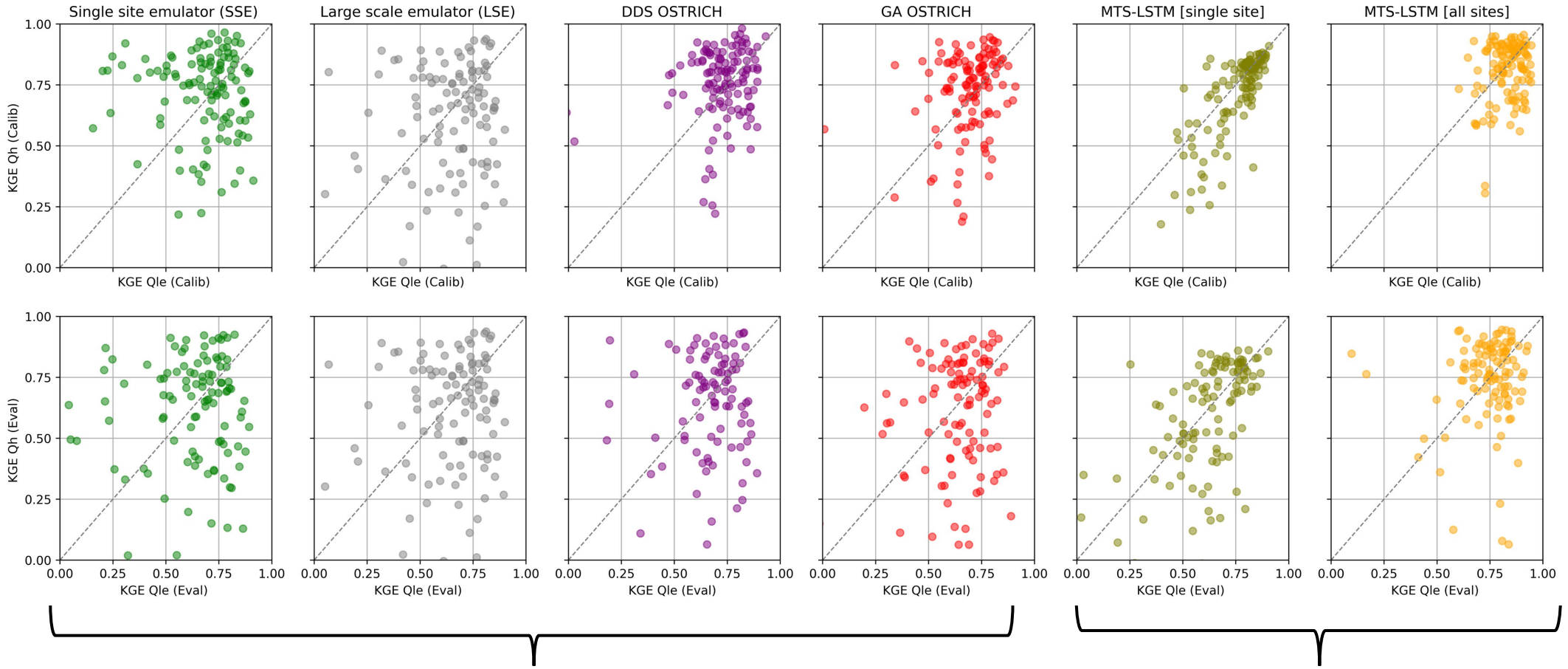
Results

The emulator methods (LSE and SSE) achieve performance comparable to traditional methods (DDS and GA) during the calibration period and outperform them during the validation period.



Results

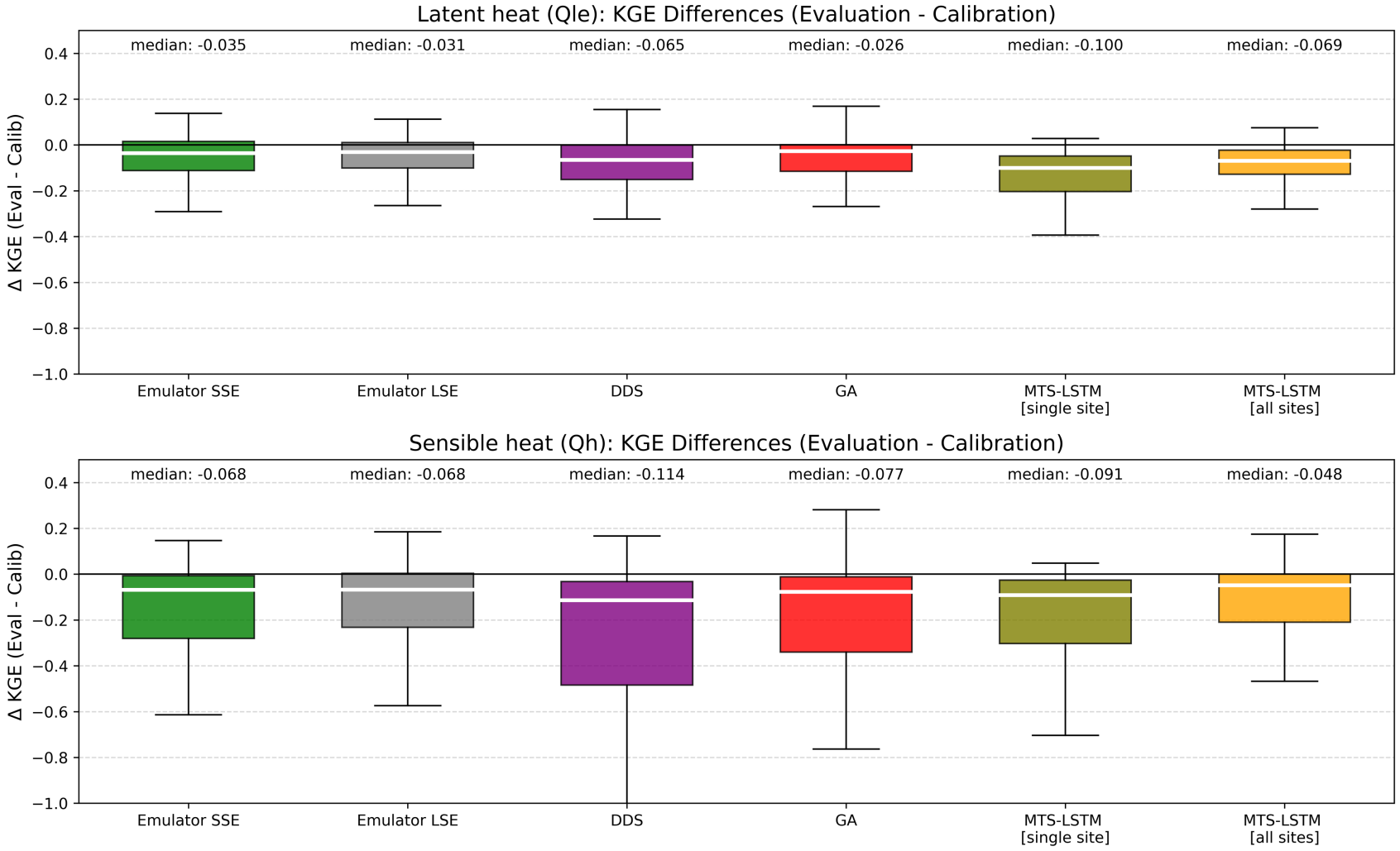
Calibration methods and benchmarks can, at the same time, improve the representation of latent heat (Qle) and sensible heat (Qh).



Results

Looking at the performance in calibration and validation periods:

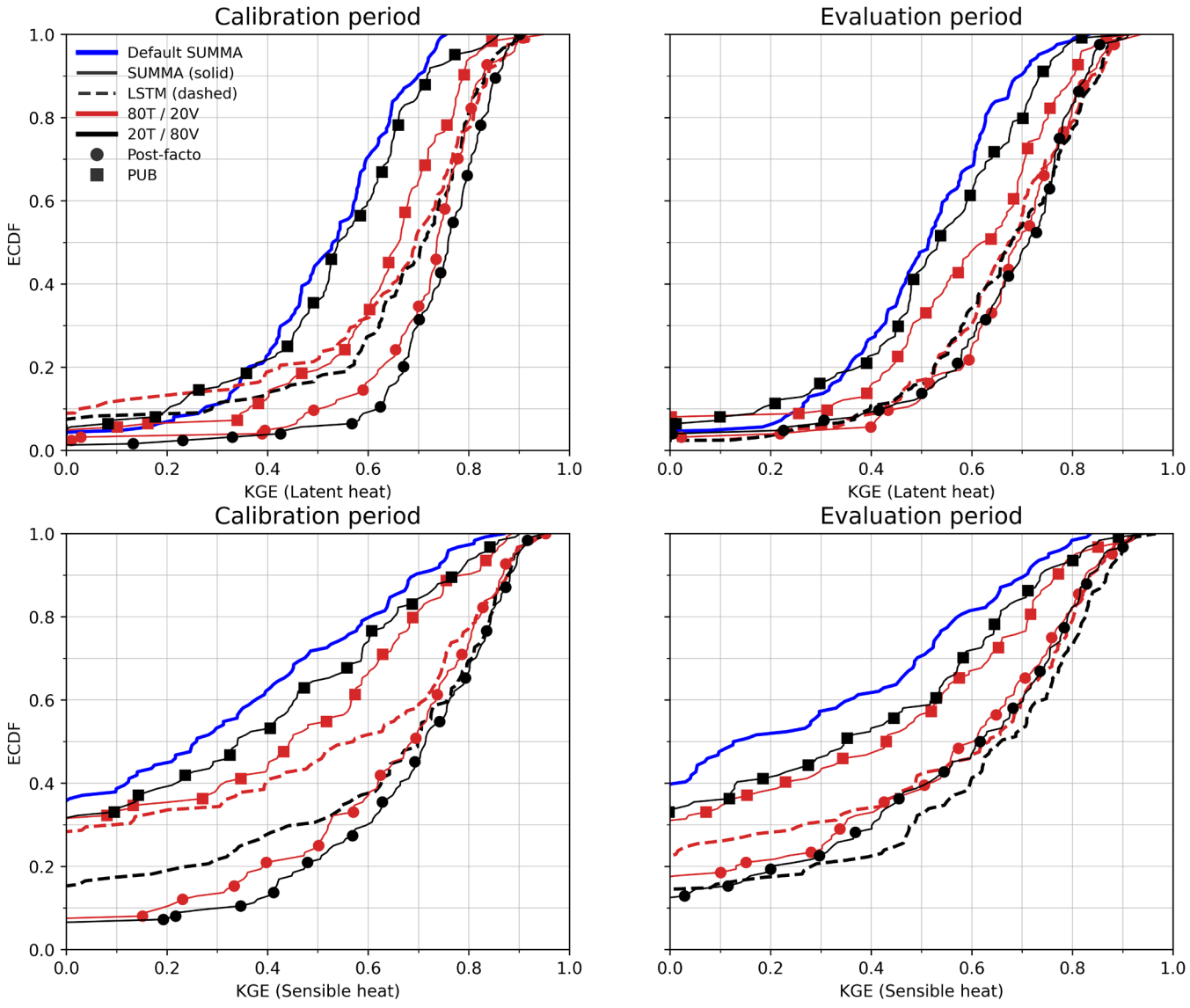
The emulator LSE can achieve the lowest overall differences between calibration and validation periods, ensuring consistent temporal validation.



Results

The emulator LSE can also be used to identify parameters in unseen flux towers, as demonstrated by the spatial cross-validation.

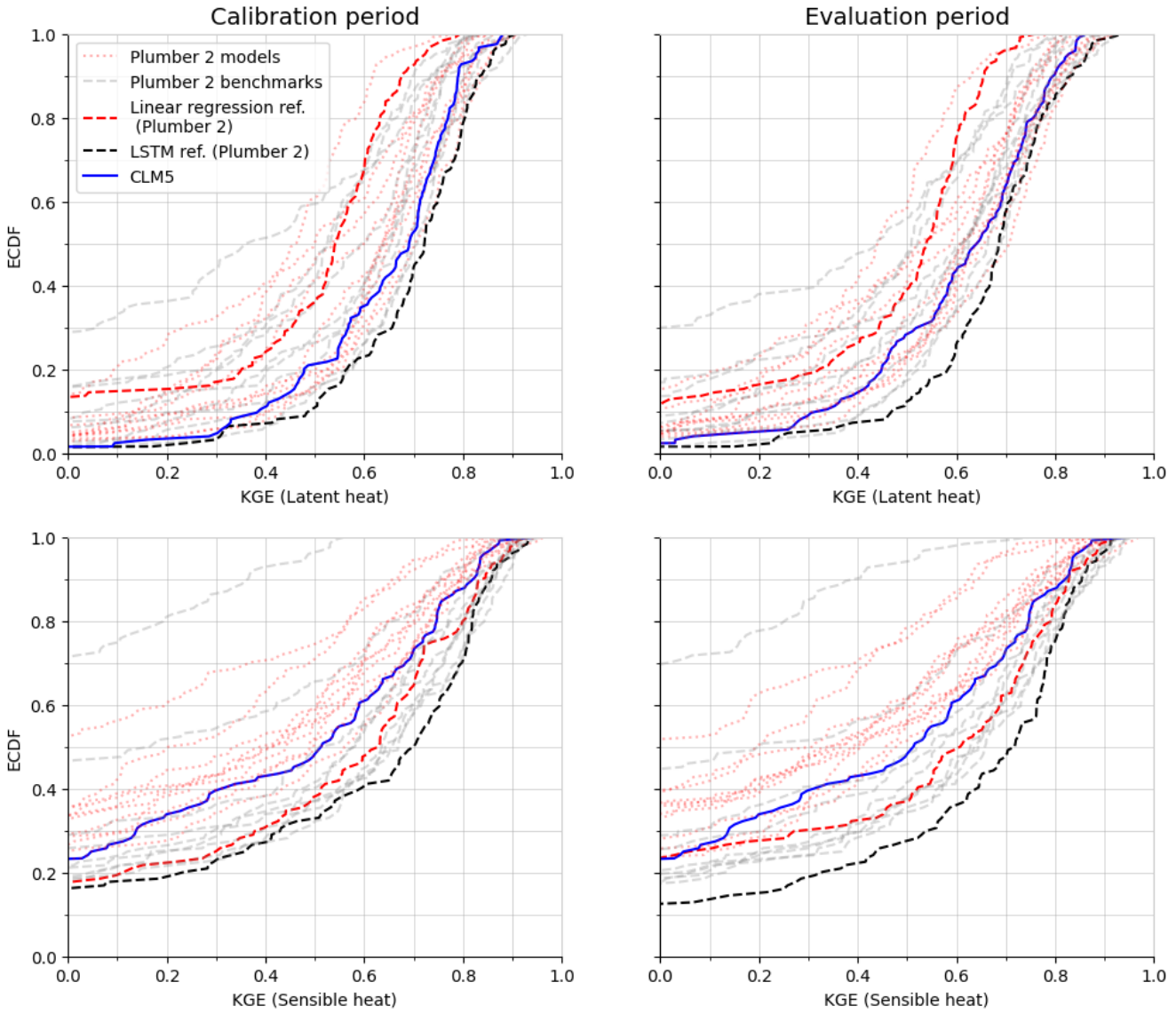
- There are results trained with 20% of the basins and tested on the remaining 80% and vice versa.
- The LSTM was built using multi-frequency MTS-LSTM (Gauch et al., 2021; Kratzert et al., 2022) using 30 epochs, a hidden size of 64, and an output dropout of 0.4.
- The LSE results can be using the simulated KGE (PUB approach) or using the actual KGE values (Post-facto)



CLM Workflow

- Using CLM and considering the demonstrated potential for performance improvement through calibration—highlighted by Tang et al. (2025), Duarte et al. (2025), Elkouk et al. (2024), and Dagon et al. (2020), among others—calibration was conducted using an emulator-based approach (Tang et al., 2025; Farahani et al., 2025).
- Previous experiments have primarily focused on the satellite phenology configuration, with fewer exploration of parameters influencing soil moisture dynamics. Accordingly, calibration was performed using the active biogeochemistry (BGC) configuration, incorporating multiple soil-related parameters.

Status



CLM Workflow

1. Identify key parameters and create 600 LHS samples.
2. Create a master case for every simulation using the ./run_tower tool developed by Teagan King et al. (2025).
3. Clone this master simulation for all samples following the protocol based on Lombardozzi et al. (2023)
    1. Spin-up with accelerated decomposition for 400 years.
    2. Normal spin-up for 100 years
    3. Check if the variables were spun up. [Northern sites required more years (~600 AD)]
    4. Run transient simulation
4. Build emulator

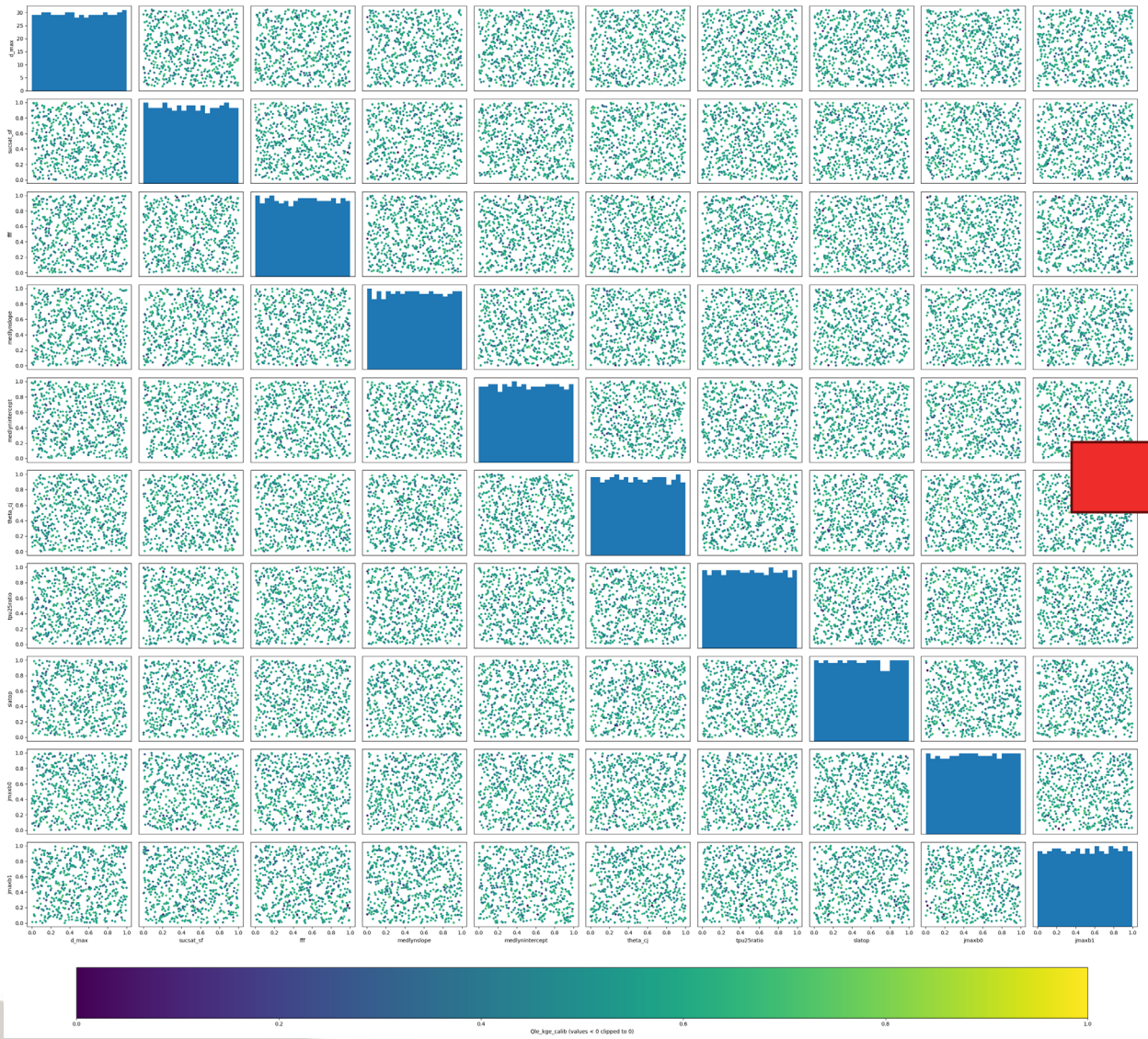
Each simulation, including the spin-up, ran for ~8–12 hours, using 1 CPU per run and up to 20GB of RAM.

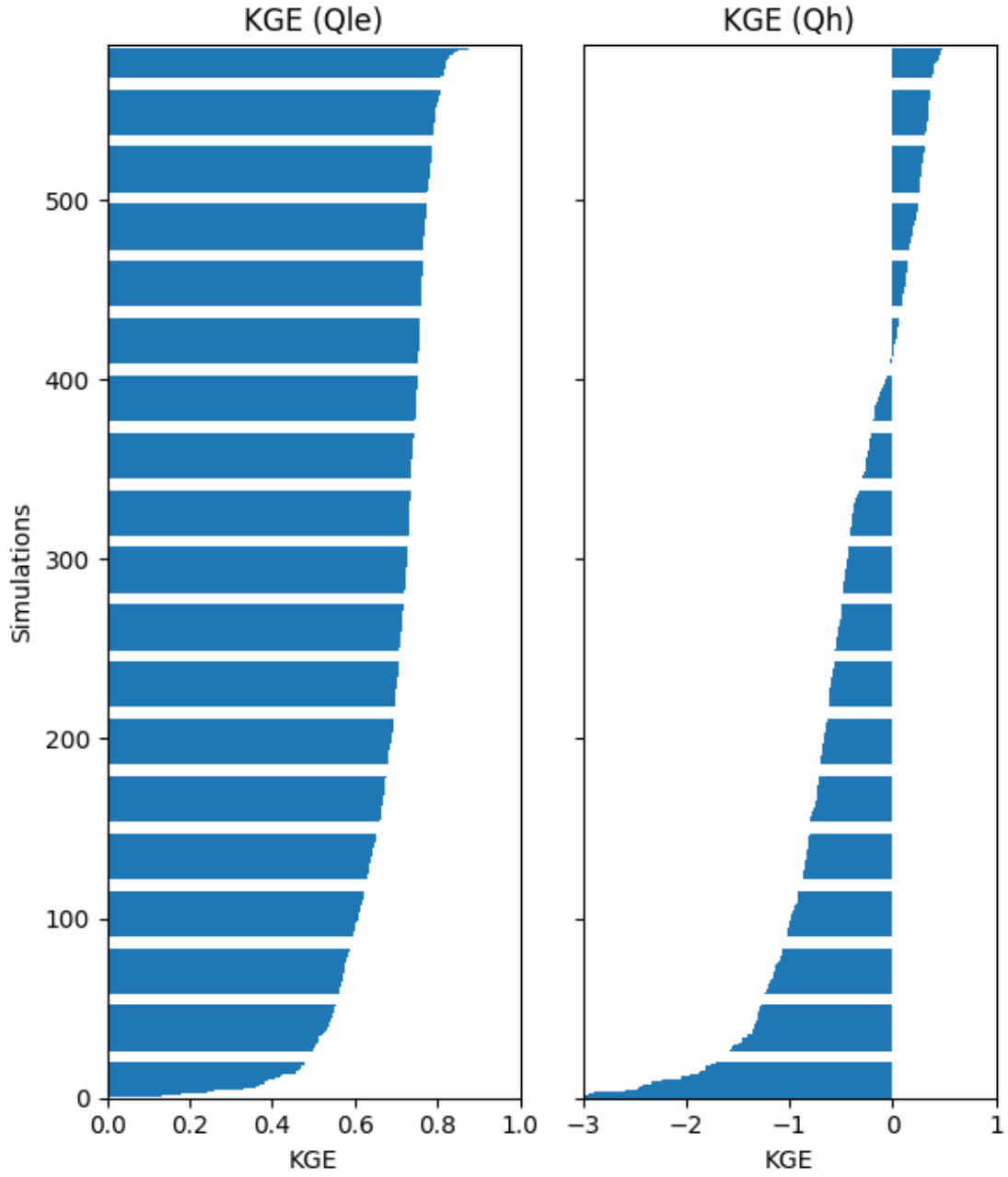
To speed up the production runs for training to the emulator, a machine-agnostic script was developed that could create the cases and either aggregate them in batches (to run on machines like Derecho that schedule entire nodes) or develop array jobs to run each individually.

Following the CWARHM's principles (Knoben et al., 2022), to keep everything simple, all simulations consulted a master control file and wrote their results (e.g., spin-up status) to a central file.

https://github.com/ESCOMP/CTSM/tree/master/python/ctsm/site_and_regional

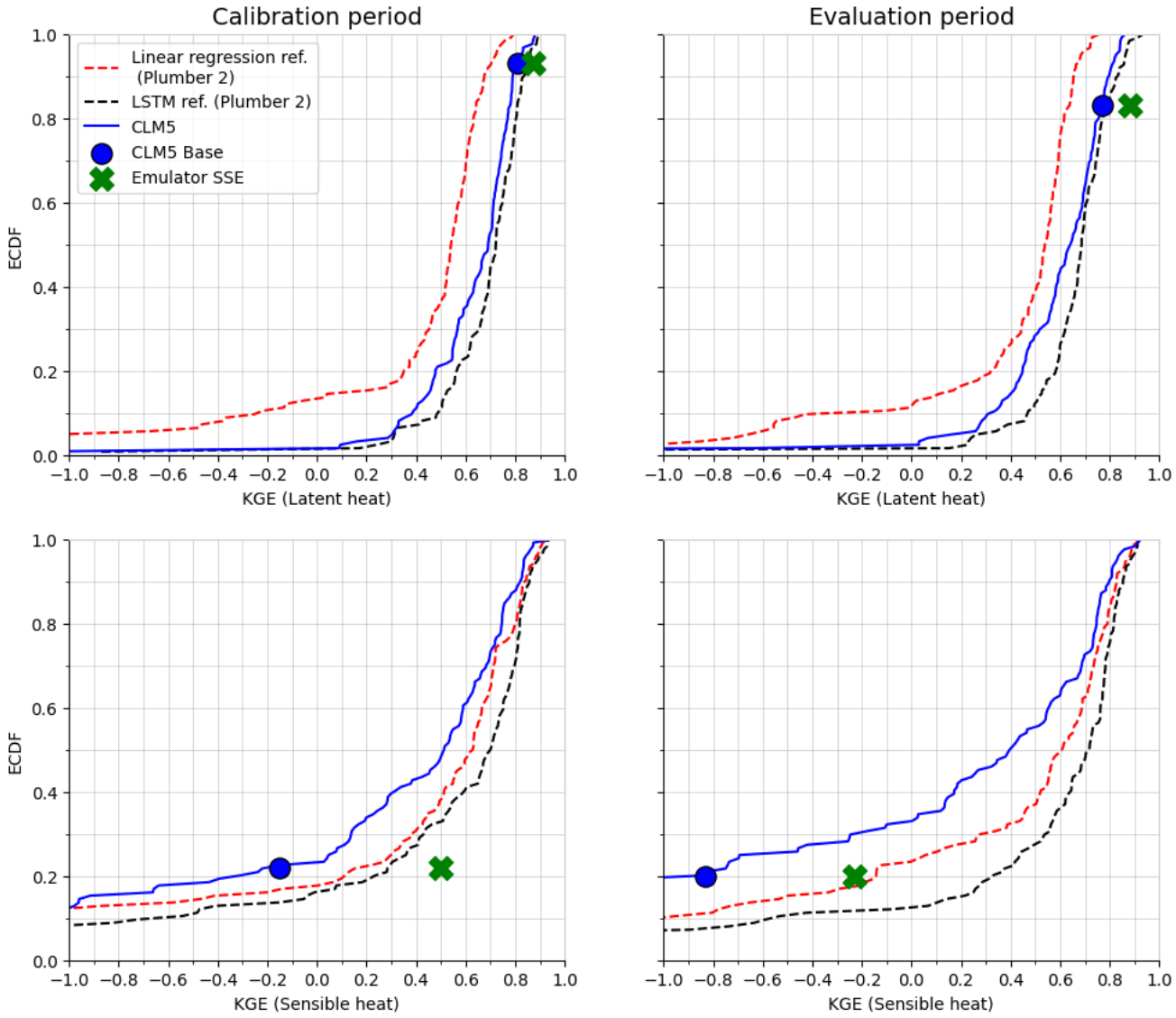
CLM results Wave 0





Preliminary results using CLM

Station AT-Neu [Austria]
Vegetation: Grasslands
Climate: Warm Summer Continental with precipitation in all seasons
Mean Annual Temp (°C): 6.5
Mean Annual Precipitation (mm): 852



Take home messages

- Calibrated results using a machine learning emulator can outperform the default parameters.
- The emulator methods (LSE and SSE) achieve performance comparable to traditional methods (DDS and GA) during the calibration period and outperform them during the validation period.
- The emulator LSE can achieve the lowest overall differences between the calibration and validation periods, ensuring consistent temporal validation.
- The emulator LSE can also be used to identify parameters in unseen flux towers, as demonstrated by the spatial cross-validation.
- The HPC-agnostic CLM workflow can improve the reproducibility of the calibration and accelerate calibration results.

Future work

UNIVERSITY OF CALGARY

- Use an emulator trained on flux tower data to identify parameter sets that best reproduce evapotranspiration patterns across North America and the planet (regionalization).
- Evaluate the impact of different model configurations (different modeling equations and parameter values), testing both default and optimized parameters to minimize errors in latent and sensible heat fluxes.

Questions?

This project received funding under award NA22NWS4320003 from NOAA Cooperative Institute Program. The statements, findings, conclusions, and recommendations are those of the author(s) and do not necessarily reflect the views of NOAA."